# TWO-STEP X-RAY TRANSIT IDENTIFICATION: BAYESIAN BLOCK SIMPLIFICATION AND SEQUENTIAL MACHINE LEARNING TECHNIQUES

## BACKGROUND

- Over 4,000 exoplanets (out of 5,000 total) were discovered in the Milky Way using the transit method within the past thirty years [1].
- The first (and currently only) extragalactic planet candidate M51-ULS-1b (from the Messier 51 galaxy) was discovered in 2021 [2].
- Such a detection shows it is likely for more planets and other potentially interesting cosmic phenomena to also be detected outside of the Milky Way.
- The use of X-ray data in such a detection indicates that Chandra may contain a rich source of data regarding the search for extragalactic planets.
- When taken together with the fact that most confirmed exoplanets in the Milky Way were also discovered with the transit method [1], this particular method is very promising in the context of future exoplanet discoveries.

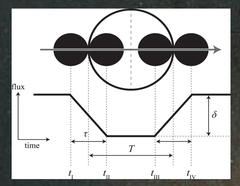


**Figure 1.** An exoplanet transit begins when the exoplanet passes between the point of observation and its star [3].

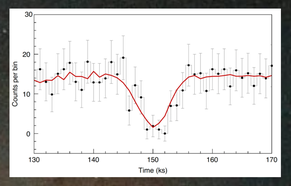


**Figure 2.** Extragalactic planet transit for M51-ULS-1b shows key characteristics: short length (~15 ks), symmetry, a significant decrease, and returning to the average value after the dip [2].

## RESEARCH PROBLEMS

- The Chandra Source Catalog contains over 36 terabytes of data from over three hundred thousand unique sources [4].
- The current process of identifying transits involves manually searching for transits in light curves (plots of the photon count rate over time).
- The immense amount of data available makes this manual process both impracticable in terms of time consumption and risky in terms of potential mistakes.

## RESEARCH OBJECTIVES

**The objective was to develop a machine learning model to accurately and automatically identify transits in X-ray data through a two-step approach, which involves the following:**

- Implementing the Bayesian blocks algorithm [5] to significantly simplify a light curve while still preserving its most important characteristics
- Generating training and validation datasets with time-series data from light curves
- Training, optimizing, improving, and evaluating a random forest model
- Testing the developed model on the validation dataset from M51, which contains a single known transit for the extragalactic planet candidate, M51-ULS-1b

## METHODOLOGY

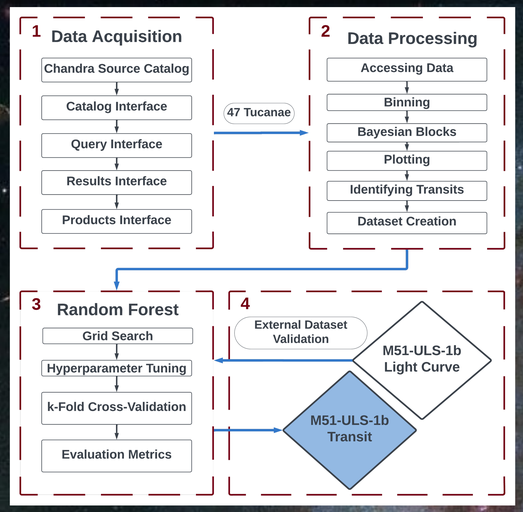


**Figure 3.** The flowchart showing the development and evaluation of the machine learning model.

### 3. Machine Learning Model

- The random forest classifier model was trained, optimized, improved, and evaluated to accurately identify transits in time-series data.
- Hyperparameter tuning to improve the model was performed using a grid search.
- The model's performance was further improved and evaluated through k-fold cross-validation.
- The evaluation metrics for each fold were calculated and then compared to one another to better understand the holistic performance of the model.

### 1. Data Acquisition

- All data was acquired from the online Chandra Source Catalog.
- The dataset from 47 Tucanae, with a significantly large number of transits, was used to train the model. The dataset from Messier 51 was used in the external validation step. (**Table 1**)

**Table 1.** Data from the two sources in 47 Tucanae and Messier 51, respectively.

| Location | Source Name | Number of Windows | | |
|---|---|---|---|---|
| | | Transits Absent | Transits Present | Total |
| 47 Tucanae | 2CXO J002400.9-720453 | 19448 | 3259 | 22707 |
| Messier 51 | 2CXO J132943.3+471134 | 1764 | 240 | 2004 |

### 2. Data Processing

- The Bayesian blocks algorithm was applied to light curves to identify short-duration dips (< 20 ks), which were then used to create the dataset to train the model.
- Changing the bin size significantly affects the complexity of the light curve. (**Figure 4**)

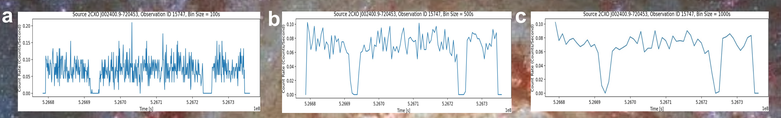


**Figure 4.** Light curves for a source in 47 Tucanae, showing the significance of bin size. The bin sizes were as follows: a) 100 s, b) 500 s, and c) 1000 s.

### 4. External Dataset Validation

- The model was tested for accuracy and robustness using data from the light curve containing the transit of the extragalactic planet candidate M51-ULS-1b.

## RESULTS

### 1. Significance of $p_0$ and Bayesian Block Simplification

- The significance value, $p_0$, affects the degree of simplification (smaller values indicate a higher threshold for significance and thus a more simplified light curve). (**Figure 5**)
- The algorithm finds the most appropriate bin size for each segment of the light curve.
- The algorithm identifies the number of edges, or points in time, that are deemed to be significant in the context of the light curve. (**Figure 6**)

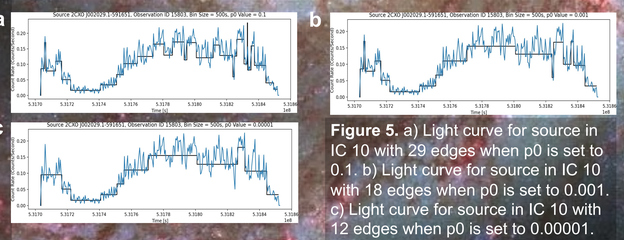


**Figure 5.** a) Light curve for source in IC 10 with 29 edges when p0 is set to 0.1. b) Light curve for source in IC 10 with 18 edges when p0 is set to 0.001. c) Light curve for source in IC 10 with 12 edges when p0 is set to 0.00001.

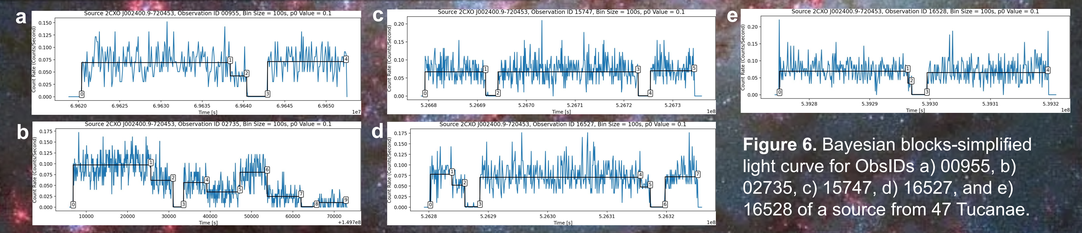


**Figure 6.** Bayesian blocks-simplified light curve for ObsIDs a) 00955, b) 02735, c) 15747, d) 16527, and e) 16528 of a source from 47 Tucanae.

### 2. k-Fold Cross-Validation

- After the most optimal hyperparameters were determined for the random forest model using a grid search, performance was further improved and evaluated through k-fold cross-validation.
- This method involves splitting the dataset into $k$ ($k$ = 6 in this work) subsets, or folds, of roughly equal size, and then training and evaluating the model $k$ times, each time using a different fold as the test set and the rest of the dataset as the training set. (**Figure 7** and **Table 2**)

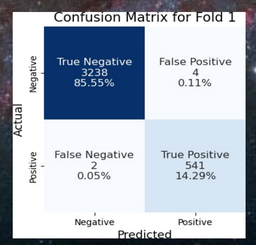


**Figure 7.** Representative image showing confusion matrix for fold 1.

## RESULTS (Cont.)

### 2. k-Fold Cross-Validation (Cont.)

- The mean and standard deviation (SD) indicate that the results for each fold are similar, showing the consistency in the model's performance. Yet, they also involve slight differences, indicating the model's ability to adapt to new data. (**Table 2**)

**Table 2.** Results from the six confusion matrices produced from the six-fold cross-validation.

| Fold | True Positive | True Negative | False Positive | False Negative |
|---|---|---|---|---|
| **1** | 541 | 3238 | 4 | 2 |
| **2** | 554 | 3226 | 0 | 5 |
| **3** | 506 | 3275 | 3 | 1 |
| **4** | 547 | 3233 | 2 | 2 |
| **5** | 550 | 3230 | 2 | 2 |
| **6** | 548 | 3234 | 1 | 1 |
| **Mean** | 541 | 3239.33 | 2 | 2.17 |
| **SD** | 17.66 | 17.92 | 1.41 | 1.47 |

### 3. External Dataset Validation

- The model accurately predicted the exact point in time of the transit of M51-ULS-1b. (**Figure 8**)

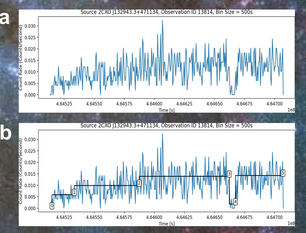


**Figure 8.** a) Light curve with bin size 500 s for M51-ULS 1b. b) Light curve with Bayesian blocks-simplified light curve, with $p_0$ significance value set at 0.1 and shown with labels. This particular light curve with this specific $p_0$ significance value has six edges.

## CONCLUSION

- A two-step process was implemented to build a successful random forest classifier model that was able to accurately identify transits in time-series data.
- Specifically, the model was quite flexible, as it was able to accurately identify transits in the galaxy M51 after training on data from the globular cluster 47 Tucanae.
- The developed model is not limited by astrophysical differences between the sources, demonstrating its robustness and applicability.
- Machine learning has not yet been applied to the specific issue of identifying transits in extragalactic X-ray data; this work shows it is possible to develop a machine learning model to accurately identify transits in such data.
- The model greatly focuses the approach to transit identification by both speeding up the process from days to minutes and reducing the possibility of human error.
- With the model, astrophysicists can perform meaningful work without needing to develop an "intuition" for transits in light curves.

## REFERENCES


- NASA exoplanet archive. 2023 [cited 2023 July 23]; Available from: https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html.
- Rosanne Di Stefano, Julia Berndtsson, Ryan Urquhart, Roberto Soria, Vinay L. Kashyap, Theron W. Carmichael, and Nia Imara, A possible planet candidate in an external galaxy detected through X-ray transit. Nature Astronomy, 2021. 5(12): p. 1297-1307.
- Imara, N. and R. Di Stefano, Searching for Exoplanets around X-Ray Binaries with Accreting White Dwarfs, Neutron Stars, and Black Holes. The Astrophysical Journal, 2018. 859(1): p. 40.
- Chandra Source Catalog Release 2.0 (CSC 2.0). [cited 2023 July 9]; Available from: https://cxc.cfa.harvard.edu/csc/.
- Jeffrey D. Scargle, Jay P. Norris, Brad Jackson, James Chiang, Studies in Astronomical Time Series Analysis. VI. Bayesian Block Representations. The Astrophysical Journal, 2013. 764(2): p. 167.


Unless otherwise noted, all graphs and tables were created by the student.